Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims

Year: 2020 | Citations: 409 | Authors: Miles Brundage, S. Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger

---

Abstract

With the recent wave of progress in artificial intelligence (AI) has come a growing awareness of the large-scale impacts of AI systems, and recognition that existing regulations and norms in industry and academia are insufficient to ensure responsible AI development. In order for AI developers to earn trust from system users, customers, civil society, governments, and other stakeholders that they are building AI responsibly, they will need to make verifiable claims to which they can be held accountable. Those outside of a given organization also need effective means of scrutinizing such claims. This report suggests various steps that different stakeholders can take to improve the verifiability of claims made about AI systems and their associated development processes, with a focus on providing evidence about the safety, security, fairness, and privacy protection of AI systems. We analyze ten mechanisms for this purpose--spanning institutions, software, and hardware--and make recommendations aimed at implementing, exploring, or improving those mechanisms.